中文情感倾向性分析的相关研究进展

吴琼 谭松波 程学旗

摘 要： 如何对大规模富含情感信息的文本进行倾向性分析是当前web应用一个亟待解决的问题。本文在分析目前国内外情感倾向性分析研究现状的基础上，介绍了我们为进行中文情感倾向性分析所构建的语料集及开发的实验平台，然后重点介绍我们的工作，包括整篇文本的倾向性分析、领域情感词典构建、跨领域情感倾向性分析等方面的关键技术，从而通过不同角度提高文本倾向性分析精度。最后总结了我们已有的工作，并展望下一步我们将深入开展的研究工作。

关键词： 倾向性分析；监督学习；情感词典；跨领域

1 引言

近年来，随着互联网在现代社会影响迅速扩大，论坛、博客等网络交流平台不断涌现，人们越来越习惯于在网上发表主观性的言论。这些言论用于表达自己对于日常事件、产品、政策等的观点和看法，形成了网上大量带有情感倾向性的文本。不同于传统的结构化的数据，这些文本的表现形式大多为非结构化或半结构化的评论文本形式。面对如此海量的富含情感信息的文本，如果仅仅依靠人工进行整理，就会面临处理周期长、费用高等问题，显然是不切实际的。因此，如何由计算机自动完成快速从大规模文本中提取出所需情感信息，进行态度分析便成为当前一项重要的研究课题。文本情感倾向性分析研究就是在这样的背景下开展起来的。

情感倾向性是一个相当广泛的概念，涉及人们的观点、看法和评价，包括人类行为相对于社会标准的评价，产品相对于国家和行业强制标准、用户偏好、审美观的评价等。文本的情感倾向包括文本所反映的情感的方向（褒或贬）及其强度。文本情感倾向性分析的目的是通过挖掘和分析文本中的立场、观点、看法、情绪、好恶等主观信息，对整篇文本所体现出的态度（或称情感倾向性），即文本中的主观信息进行判断。文本情感通常分为两类(正面、反面)或三类(正面、反面和中立)。其中正面类别(positive)是指主题中持有积极的(支持的、健康的)态度和立场；负面类别(negative)是指文本中持有消极的(反对的、不健康的)态度和立场；中立类别(neutral)是指文本中持中立态度和立场。从当前的研究来看，以考虑两类的研究居多。

文本倾向性分析与传统的文本分类不同。传统的文本分类基于文本主题（例如：文化、体育、经济等）进行分类，对文本内容的分析与理解都处于比较浅的层次。而文本倾向性分析关注的是非主题分析，即文本内容所体现的情感、态度，而非文本本身的内容。它是对传统的文本分类研究的深入和拓展，可以满足人们更深层次获取和利用信息的要求[1-3,21-26]。

网上文本的形式及内容的随意性使文本情感倾向性分析具有很高难度，涉及到人工智能、机器学习、信息抽取、信息检索、数据挖掘、自然语言处理、计算语言学、语料库语言学、本体学(ontology)、统计学等多个研究领域，不仅需要应用上述领域前沿技术，而且反过来又对这些领域提出了新的挑战，推动了其发展。因此，在科学研究方面具有重要意义。同时，文本倾向性分析可广泛应用于社会舆情分析、产品在线跟踪与质量评价、影视评价、博客声誉评价、新闻报道评述、事件分析、股票评论、图书推荐、企业情报系统、客户关系管理(CRM)等方面，在社会经济和人民生活方面也具有重要意义[1]，如：

- 社会舆情分析：舆情是指在一定社会空间内，围绕社会事件的发生、发展和变化，民众对事件和当事各方的社会政治态度，是人们对于社会中各种现象、问题所表现的信念、态度、意见和情绪等总和。网络因其开放性和虚拟性，已经成为民意表达的重要通道和空间。利用文本倾向性分析技术，可以更加及时地了解网络民意，使民间智慧与官方智慧更加良好地互动。

- 博客声誉评价及垃圾博客过滤：及时的交互性是博客的特色之一。大量网民利用博客发表自己对事物的观点并对他人的观点进行评论，博客浏览者也大多根据评论信息来判断博客作者的声誉，与浏览者的互动是很多博客作者继续写博客的动力。利用文本倾向性分析技术可以挖掘浏览者对博客作者的褒贬观点，从而得到博客作者的声誉度。此外，可以通过倾向性分析技术对以广告等垃圾信息为主的博客进行过滤。

- 产品评价与推荐：目前，多数产品生产、销售厂商希望通过跟踪用户对产品的回馈意见来获得改进产品质量的针对性意见；潜在的消费者也希望通过网上真实的评价信息来调整个人的购买意向。然而，随着评论数量的快速增长，商家和潜在消费者都希望能有一种方法来帮助他们自动对这些产品评论进行处理。利用文本倾向性分析技术对产品评论观点进行组织和分类，有利于人们了解产品，培育潜在消费群体。

- 影视评价：影视评价是影视艺术与观众的桥梁，是实现影视作品三重价值(艺术、社会、经济)的重要手段。影视评论主要是对影视作品的主题、拍摄、情节、人物形象、人物语言风格、表演技巧、画面等方面进行分析，同时，就影视作品的音乐设计、画面特点、人物服装及化妆造型、人物与环境的搭配、色彩使用等方面发表见解进行评价。文本情感倾向性分析技术可以实现影视评论的自动分类，有利于用户快速浏览正反两方面的评论意见，减少观看影视时的盲目性。

综上所述，文本倾向性分析研究不但具有深远的理论价值，而且有着广阔的应用前景，可以创造巨大的社会和经济效益。

本文针对现有的文本倾向性分析方法所存在的问题，在分析目前国内外倾向性分析研究现状的基础上，介绍了我们所构建的语料集及开发的实验平台，然后重点介绍我们的工作，包括整篇文本的倾向性分析、领域情感词典构建、跨领域情感倾向性分析等方面的关键技术，从而通过不同角度提高文本倾向性分析精度。最后总结已有的工作，并展望下一步将深入开展的研究工作。

本文第 2 节将概述国内外研究现状；第 3 节介绍我们的语料集及实验平台；随后分 3 节详细介绍我们在情感倾向性分析方面的研究工作：基于监督学习的情感倾向性分析研究、领域情感词典构建以及跨领域情感倾向性分析；最后在第 7 节总结我们的工作并展望下一步的研究。

2 国内外研究现状

文本倾向性分析研究的历史不长，最早可以追溯到 20 世纪 90 年代，并且在 2000 年之后获得了突飞猛进的飞速发展。目前，文本倾向性分析研究已成为国内外研究的热点问题。近年来，有关自然语言处理、人工智能、信息检索、数据挖掘以及 Web 应用等领域的多个国际顶级会议(AAAI、ACL、CIKM、COLING、SIGIR、WWW 和 KDD 等)涌现出大量文本情感倾向分析的相关论文。同时也出现了针对文本倾向性分析的相关评测，例如，从 1992

年起，美国国家标准技术研究院（NIST）和美国国防高级研究计划署（DARPA）组织了文本检索会议（TREC），该会议已经成为国际上文本检索领域最著名的评测会议；2006 年起，TREC 增加了博客观点（Blog Opinion）检索任务，在全球范围内开展博客倾向性观点的检索和分析研究。在政府方面，美国的舆情研究协会、欧盟舆情分析官方网站、新西兰坎特伯雷大学（The University of Canterbury）欧洲舆情分析研究中心等都开展了基于调查问卷、网页统计、文本分析等方式的舆情倾向性分析项目。

下面，我们首先分别对文本情感倾向性分析的国内外相关研究分类进行综述，在此基础上归纳主要应用的分类技术。

2.1 代表性工作

文本情感倾向性分析根据其所处理的情感数据粒度不同分为：属性级的倾向性分析、词语级的倾向性分析、文档级的倾向性分析以及对于多文档的倾向性摘要等[3,14,15,17]。

（1） **属性级的倾向性分析**

属性级的倾向性分析针对细粒度的文本挖掘，主要包括两方面的研究内容：评论语气词识别、评论对象的识别以及其与评论语气词的关联[4,5]。

（2） **词语级的倾向性分析**

词语语义倾向计算是文本倾向性分析研究中的一个基础且重要的子研究领域，其目标是提供文本倾向性的量化表达。即用(-1,1)之间的实数代表词语的语义倾向，其正、负分别代表语气的褒、贬，绝对值代表词语的极性强度，这为文本倾向性分析的多个研究方向提供了重要基础[1]。目前，词语语义倾向性分析除了利用预先标注的语义倾向基准词外，还需要利用词语间的相似度[6-10]。

（3） **文档级的倾向性分析**

文档级的倾向性分析可以看作是一种特殊的分类，即根据文章中对某一主题的观点（支持或反对、高兴或悲伤等等）对文本进行分类，因此可将机器学习算法用于这种分析[11,12]。

（4） **多文档的倾向性摘要**

目前网上包含主观信息的文本中，在线的产品评论，尤其是针对某些名牌产品的文本数量增长极快。多数产品评论篇幅较长，但包含产品属性的句子却极少。对于潜在消费者来说，难于在如此海量的信息中找到真正有价值的评论。而对于产品生产、销售商来说，在如此众多的评论信息中跟踪消费者对于自家产品的评价也是一件相当困难的事情。

因此，产品评论挖掘系统通常也要利用意见摘要技术，通过归纳评论的语气极性、程度和相关事件对在线产品评论进行摘要。利用该技术，潜在用户可以方便地了解目前消费者对于产品的评价；产品生产、销售商也可以较轻松地跟踪消费者对于产品的评价，比较同类各品牌产品的优劣[13]。

2.2 主要的分类方法

目前常用的文本情感倾向性分析技术主要有：统计机器学习方法、基于相似度的方法、基于图模型的方法。

2.2.1 统计机器学习方法

当前，基于统计机器学习理论的文本情感倾向性分析是文本挖掘领域的一个研究热点。其中常用的基于机器学习的文本分类算法包括 [1,3]：

- **中心向量分类方法**：这是一种简单有效的分类法，所有文档都用特征向量来表示。在此基础上，对于所有属于同一类别的文档计算出一个平均向量（即中心向量）。给一个样本向量分类时，只需计算它与各中心向量的相似度，取相似度最大值的中心向量所在类别作为样本的类别即可。

- **k-近邻**（K-Nearest-Neighbor，**KNN**）**分类方法**：这是一种非常有效的归纳推理方法，直观地讲，k-近邻分类方法就是从测试文档 d 开始生长，并不断扩大区域，直到包含 k 个训练样本点为止，并且把测试文档 d 的类别归为这最近的 k 个训练样本点中出现频率最大的类别。

- **贝叶斯分类器**：朴素贝叶斯分类器是一种通用的监督学习算法。该方法首先将已标注倾向性的文本作为训练样本，并选取句子中的单词及词性标签等作为分类特征。另外，语句中语气词出现的数量也被当作判定文本倾向性的一个依据，然后将这些特征作为输入，利用贝叶斯公式对待标注文本进行分类。

- **支持向量机**：这是传统分类中非常有效的一种方法，它的分类结果比朴素贝叶斯方法普遍要好。其基本思路是：给定一个训练集，找到一个具有最大间隔的分隔平面（也称超平面）$\vec{\omega}$，作为类别的分界。间隔越大，得到的分类器也越好。基于文档特征向量，通过语气挖掘将文档分为正面和负面两类。采用支持向量机方法相当于求解一个带约束条件的最优化问题。

- **条件随机场**：这是一个在给定输入节点（也就是观察值）条件下计算输出节点（也就是标签）的条件概率的无向图模型。条件随机场模型特别适合处理序列标记问题，在属性级的情感倾向性分析研究中，被应用于标记评论语气词与评论对象的关联。

- **最大熵分类器**：这是一种通用的监督学习算法。利用该技术可以将主观性文本和客观性文本分开。该算法的思想是为所有已知的因素建立模型，而把所有未知的因素排除在外。也就是说，要找到这样一个概率分布，使其满足所有已知的事实，且不受任何未知因素的影响。该算法首先将已标注倾向性的文本作为训练样本，从中抽取出单词、词性标签等作为特征，另外语句中语气词出现的数量也被当作判定文本主观性的一个依据。然后利用这些特征和最大熵模型为待标注文本判定倾向性。

2.2.2 基于相似度的方法

基于相似度的方法的基本思想与 K-近邻方法类似，即利用 K 个已标记的样本点，通过样本之间的相似度，来对新的样本进行标记。基于相似度的方法采用语句间公共单词、短语的数量以及语义词典中的词语相似度来计算语句的语义相似度[9]。

2.2.3 基于图模型的方法

对于倾向性分析问题，可以利用词语或文本语义关系构建图，将词语或文本看作图中的顶点，利用词语间或文本间的关系为图增加连边，形成一个图模型，然后根据此模型及其相应算法来进行倾向性分析。大量的研究人员基于图模型的方法进行了研究，产生出一系列成果[9]。

3 实验设计和实验平台

3.1 数据集

倾向性分析研究离不开数据集。然而，目前倾向性分析研究尚处于初级阶段，国际上仅有一两个公布的小规模语料集，而国内研究则处于起步阶段，尚未见到公开的可用语料集。因此，构建一定规模的标准语料集是进行倾向性分析的必要基础。我们参照类似美国语言数据联盟(Linguistic Data Consortium，LDC)以及路透社、文本检索年会(Text REtrieval Conference, TREC)[1]、话题监测与追踪研究(Topic Detection and Tracking，TDT)[2]等国际机构建立的评测数据集的标准，采用自主研发的大规模网络信息采集技术获取网络评论文本，通过自动和人工标注相结合的方法，建立具有一定规模的文本倾向性分析标准数据集。在此基础上开展有效的倾向性分析算法的研究。

目前我们已经从互联网的相关中文评论网站采集并标注了影视、教育、房产、笔记本电脑（简称电脑）、手机、电子产品（简称电子）、股票、酒店以及书籍九个主题的中文评论数据将近 17000 条。由于同一主题的评论可能出现在不同的评论网站，为防止数据集中出现重复的样本，对于特定的网页地址我们指定了特定的采集者。语料采集后，经过抽取，转换成统一的文本格式，经过自动标注和人工校对文本极性（正面评论或负面评论），最终得到可用的数据集。数据集中的样本情况如表 1 所示：

表1. 倾向性分析数据集样本情况

主题	样 本 数		
	总数	负面	正面
影视	1980	1062	918
教育	1476	1012	464
房产	1118	733	385
电脑	901	451	450
手机	992	497	495
电子	1608	554	1054
股票	1047	683	364
酒店	4000	2000	2000
书籍	4000	2000	2000
合计	17122	8992	8130

数据集中教育、房产、电子三个主题的正面和负面评论数量上存在较大的差异，其它主题的正面和负面评论数量相当。各类评论文档的长度各异。影视类评论的平均篇幅最长，约为 500 个汉字;手机类评论的平均篇幅最短，约为 60 个汉字。

3.2 评价标准

情感倾向性分析（包括情感词典构建）研究通常使用四种标准评价：准确率（Precision），召回率（Recall），F 值和精度（Accuracy）。

设 a_1 表示分类器判断为正向，且与人工标注结果一致的样本数； a_2 表示分类器判断为负向，且与人工标注结果一致的样本数； b_1 表示分类器判断为正向的样本数； b_2 表示分类

[1] 由美国国家标准研究院（NIST）组织
[2] 由美国语言数据联盟组织

器判断为负向的样本数；c_1表示人工标注为正向的样本数；c_2表示人工标注为负向的样本数，则准确率计算公式为：

$$Precision = \frac{a_1 + a_2}{b_1 + b_2} \qquad (1)$$

召回率计算公式为：

$$Recall = \frac{a_1 + a_2}{c_1 + c_2} \qquad (2)$$

F 值计算公式为：

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (3)$$

在许多实际 Web 应用中，通常会为了提高准确率而适当牺牲一些召回率。

此外，还可以使用精度作为倾向性分析系统的评价标准，其定义如下：

$$Accuracy = \frac{\text{分类正确的文本数}}{\text{测试文本总数}} \qquad (4)$$

3.3 实验平台

为了使用户能够方便、快捷、自动进行文本倾向性分析，我们开发了一套文本倾向性分析系统。该系统可以对采集到的文本进行倾向性分析，并给出最终判别结果。该系统的原理下文详述，系统界面如图 1 所示。







图1 倾向性分析系统界面

4 基于监督学习的情感倾向性分析

基于监督学习的文本倾向性分析是当前的研究热点。然而，基于监督学习的倾向性分析方法还有许多亟待解决的问题：（1）如何确定各种有监督的学习方法在中文数据集上的倾向性分析效果孰优孰劣；（2）文本特征表示方法和特征选择机制等因素对中文倾向性分析的性能将产生什么影响；（3）文档集的哪些情感特征对倾向性分析的精度具有决定性影响，等等。本节集中研究前两个问题，通过分析常规分类方法的特点，研究各种特征表示和特征选择方法对倾向性分析结果产生的影响，并对实验结果进行了详细对比分析[3]。

4.1 基本原理

基于监督学习的方法主要包括以下几个方面的内容：倾向性特征的提取、表示、压缩和倾向性分类器的训练[2]。

（1） 特征提取

对于一个原始文档，将其切分成句子，然后通过语法分析器有选择地进行词语切分，经过一系列转换后形成向量列表，这个过程称为特征提取。有很多处理方法可以应用于特征提取：

词典过滤方法：可分为基于 WordNet 的方法和基于语法成分（part-of-speech，POS)标注的方法两种。WordNet 过滤法是将词语替换成 WordNet 中可能的同义词集合。POS 标注法认为，对于单个短语或词汇来说，只有当它们在句子中充当某些特定的成分时，才可能用来表达观点倾向性，充当其它成分时，就属于倾向性分析的噪音，应当被过滤掉。

形容词评价方法：立足于提取和分析形容词评价组，这些形容词评价组由一个主干形容词（如“beautiful” 或 “boring”）通过有选择地与一系列修饰成分（如“very”, “sort of”, or “not”）组合而衍生出来。

（2） 特征表示

进行倾向性分析之前，首先需要把数据集中的文本表示成特征，这可以采用反映文本语言学特征的元素来表示，如使用词、n-Gram[3]、词组和概念等。向量空间表示模型（Vector space model，VSM）是目前文本表示的主要方法，相关研究集中在以什么语义单元作为项及如何计算项的权重两个问题上，通常以项的出现频率作为基础计算权重。也有一些文本表示方法希望通过借鉴自然语言处理技术，考虑被词袋忽略的语义单元间的联系，将词义及短语等复杂的项应用到分类方法的文本表示中。不过这些表示方法在分类效果上还没有明显的优势，而且往往需要比较复杂的语言预处理，在分类时会影响分类器的速度。到目前为止，非 VSM 的表示在理论上的合理性及面对实际应用的可扩展性还需要深入验证。

（3） 特征选择

用特征表示方法生成的特征中可能存在很多噪声，通过特征选择舍弃一些不太重要的特征，将有效消除噪声的影响，降低向量空间的维数，简化计算，防止过分拟合。特征选择是根据某种准则从原始特征中选择部分最有类别区分能力的特征。常见的用于特征选择的衡量标准有文档频率、信息增益、互信息和 CHI 统计等，因其复杂度较低而应用广泛。特征数量的变化和分类器效果紧密相关。有关文献的结论表明：合理的特征选择方法会使多数分类器的性能快速提高并能迅速接近平稳；但若特征数目过大，分类器的性能反而可能出现缓慢降低。

（4） 文本分类器选择

有很多文本分类方法可应用于倾向性分析，如中心向量法、k-近邻法、感知器分类法（Winnow 算法）、朴素贝叶斯法和支持向量机方法等。

4.2 实验及结果分析

我们采用影视、教育、房产、电脑和手机五个主题的中文评论数据进行了下述实验。

（1） **基于 n-Gram 的特征表示方法实验结果分析**

本实验中，我们分别采用了三种类型的特征表示方法，即 UniGrams、BiGrams 和

[3] N-Gram 是大词汇连续语音识别中常用的一种语言模型，一个“n-gram” 就是在给定序列中的一个包含 n 项的子序列

TriGrams。其它实验条件相同，即对于每个主题，使用 50%的数据作为训练集，剩余 50%的数据作为测试集，选取全部特征，使用支持向量机分类方法。实验结果如下：

表2. 基于 n-Gram 的特征表示方法分类精度比较

	影视	教育	房产	电脑	手机
UniGrams	83.0	97.9	96.1	92.0	97.4
BiGrams	83.0	97.4	96.8	94.2	97.2
TriGrams	81.8	93.1	92.7	91.6	96.6

从上表中可以看出，整体而言，BiGrams 要略好于另外两种。

（2） **基于不同词性的特征表示方法实验结果分析**

通过对评论语料进行分析，我们发现，倾向性分析与其它分类的差别在于，情感的正面表达和负面表达主要以形容词、副词和少数动词和名词的表达为主。因此，我们使用不同词性的词来表示特征，对选取四种词性中的一种和选取它们的全部（下表中的 nvaa）分别进行了实验，实验结果如下：

表3. 不同词性作为特征表示方法的分类精度比较

	影视	教育	房产	电脑	手机
名词	78.2	95.4	95.7	69.4	88.1
动词	71.1	94.9	94.5	73.9	84.1
形容词	63.0	86.2	79.2	74.9	82.2
副词	58.6	86.2	69.8	73.6	80.2
nvaa	81.5	97.4	96.4	89.8	96.8

从整体实验结果来看，以单个词性为特征的分类精度均比表 2 中 n-Gram 为特征的分类精度要差很多。以四种词性为特征的分类精度却能和 n-Gram 的精度相当。对于单个词性而言，各领域中基本都是名词和动词作为特征的分类精度要比形容词和副词的结果好，只有个别领域有不同情况。这与预想中形容词和副词带有绝大部分情感特征的想法有较大差异。

（3） **基于不同的特征选择方法实验结果分析**

我们分别采用了互信息(MI)、信息增益(IG)、CHI 统计量(CHI)和文档频次(DF)四种不同的特征选择方法进行了实验。其它实验条件相同，实验结果如表 4 所示。从中可见，使用互信息和 CHI 进行特征选择，由于它们对低频词的倚重，必定会将更多的低频词作为特征使用，从而导致了分类效果不如文档频次。而信息增益不但考虑了类别信息，而且考虑了低频词对分类结果的影响，因此分类效果最好。

（4） **基于不同分类方法实验结果分析**

表4. 不同特征选择方法的分类精度比较

	互信息	信息增益	CHI	文档频次
影视	61.2	75.7	62.9	67.5
教育	83.3	97.2	83.7	92.4
房产	73.1	96.6	93.7	93.7
电脑	72.1	93.3	90.2	91.6
手机	63.7	97.3	94.2	95.8

表5. 不同分类方法的分类精度比较

	中心向量	K-近邻	感知器	NB	支持向量机
影视	79.5	80.4	74.6	78.2	83.0
教育	95.0	96.6	93.2	95.8	97.4
房产	94.0	95.5	88.7	95.2	96.8
电脑	92.5	88.9	82.0	89.4	94.2
手机	95.8	92.1	87.9	96.2	97.2

以 BiGrams 作为特征表示方法，对于每个主题，使用 50%的数据作为训练集，剩余 50%的数据作为测试集，在选取全部特征的情况下，我们分别采用了中心向量、K-近邻、感知器、

朴素贝叶斯和支持向量机五种不同的分类方法进行了实验。实验结果见表 4 和表 5。

在以上几种分类方法中，相比而言，支持向量机的分类效率较低，但精度明显高于其它方法。

（5） 基于不同特征数量实验结果分析

采用信息增益选择的特征，按照特征权重值大小降序排列，选取权重值靠前的一定数量（500 个, 1000 个, …, 10000 个）特征进行实验，精度随特征数量变化的情况如表 6 所示：

表6. 不同数量特征的分类精度比较

特征数量	500	1000	2000	3000	4000	6000	8000	10000
影视	62.8	65.8	71.3	73.6	75.8	73.3	75.3	75.7
教育	94.2	94.7	96.5	97.2	97.3	97.7	98.1	97.2
房产	90.5	91.8	93.4	95.2	96.1	95.0	95.9	96.6
电脑	86.0	88.9	88.5	91.6	92.0	93.3	93.8	93.3
手机	93.8	95.6	96.8	94.8	95.8	96.0	97.6	97.4

从实验结果来看，对于一定的分类数据集，并非选择的特征数量越多越好。

（6） 基于不同规模训练集的实验结果分析

表7. 不同规模训练集的分类精度比较

规模	全部	1/2	1/3	1/4	1/5	1/6	1/7	1/8	1/9	1/10
影视	83.0	80.0	78.1	73.7	75.7	71.6	68.8	70.1	69.4	69.6
教育	97.4	95.7	93.5	92.0	90.4	87.9	85.9	82.8	82.8	82.4
房产	96.8	95.5	94.3	91.6	90.2	88.6	82.8	80.9	84.6	77.8
电脑	94.2	90.2	87.4	84.9	84.5	81.4	84.3	80.3	80.9	79.8
手机	97.2	96.6	94.8	93.5	95.2	86.9	93.0	91.0	91.6	91.6

前面各实验中，对于每个领域，都使用了 50%的数据作为训练集，剩余 50%的数据作为测试集。表 7 的实验用来考察不同规模的训练集对分类精度的影响，我们分别选取了训练集的全部、1/2、1/3…直到 1/10，在全部测试集上进行实验，结果如表 7 所示。说明通常情况下，足够大的训练集对于较高的分类精度具有决定性作用。

4.3 本节小结

通过本节实验表明：在基于监督学习的中文倾向性分析中，语料的语言风格对分类结果有一定的影响；文档频次特征选择方法相对于互信息、CHI 和信息增益等特征选择方法并不占优；n-Gram 特征表示方法能产生良好的结果，而单一词性的词并不能反映网络评论的整体情感特征；整合更多情感表达的特征来进行分类，才可能提高分类精度；和其它分类方法比较，支持向量机分类方法在精度方面具有明显的优势；在数据集一定的情况下，特征空间的维数并非越多越好，分类精度将在一定的维数达到最大值；通常情况下，足够大的训练集对于较高的分类精度具有决定性作用。总之，采用 n-Gram 特征表示方法、信息增益特征选择方法和支持向量机分类方法，在足够大训练集和选择适当数量特征的情况下，倾向性分析能取得较好的效果[7]。

5 领域情感词典构建

对一篇文档而言,能对其语义倾向起到决定性作用的主要是构成这篇文档所用的词语。所以,对文本进行基于情感的文本分类的基础是判定词的语义倾向。但是，目前不论是英语还是汉语，都没有，也不可能有一个完整的涵盖所有词语的语义倾向词典，因为很多的词语在不同语境中的语义倾向不尽相同。因此，对于文本倾向性分析研究来说，设计高效的情感词典构建算法是一个相当基础而且重要的工作，对于推动文本倾向性分析技术的发展、发挥文本倾向性分析的潜力并促进其实用化和商业化具有重要的现实意义。

本节以通用情感词典构建及领域情感词典构建为目标，从以下几个方面来研究该问题[23-25]。

5.1 基于函数优化的通用情感词典构建

本节提出一个可扩展的词汇语义倾向计算框架,将词语语义倾向计算问题归结为优化问题。在算法实现上，首先利用多种词语相似度计算方法构建词语无向图，然后利用以“最小切分”为目标的目标函数对该图进行划分，并利用模拟退火算法进行求解[1]。

5.1.1 基本原理

假定用一个无向图来表示字典中所有词语的关系。本文基于这样的假设：具有较大的相似度的两个词语更有可能具有相同的语义倾向。这样，词语的语义倾向计算问题可以归结为对图进行划分，使得符号相同的节点子图相似度之和最大；同时，让符号相异的节点子图相似度之和最小。这样，就确定了图中每个词语的语义倾向。

本文以“最小切分”为目标对图进行划分，目标函数需满足以下几个条件（1）奖励子类内部的连边；（2）惩罚子类内部的非连边；（3）惩罚子类间的连边；（4）奖励子类间的非连边。同时，可以将目标函数所满足的条件归为两类：条件（1）和条件（2）用于增加子类的内聚性；条件（3）和条件（4）用于减少子类之间的耦合性。

这样，我们得到了一个可扩展性较好的词语语义倾向计算框架：

（1） 利用词语间关系构建词语无向网络图（本文分别使用基于词典和基于语料两种方法）。

（2） 将词语语义倾向计算问题转化为图划分问题，并进一步转化为函数优化问题（本文以“最小切分”思想设计目标函数）。

（3） 构建求解算法对目标函数进行求解（本文使用模拟退火算法进行求解）。

在下面小节中，将分别介绍该框架的几个组成部分。

5.1.2 词汇相似度计算

词汇相似度是用于度量词语之间的相似程度。通常，相似度值被定义为 0 到 1 之间的一个实数，绝对值越大，相似度越高。本文分别采用了基于语料统计的相似度计算方法和《知网（HowNet）》提供的词语相似度计算方法作为构建词语无向网络图的基础。

（1） 基于共现率的词语相似度

互联网作为一个巨大的语料库，其价值已被越来越多的人认识。可以利用搜索引擎，将传统的基于词语共现率计算相似度的方法进行适当变化，使其可以应用于互联网语料。利用上述各种方法可以得到的两两词语之间的相似度，构造词语无向网络图。

（2） 基于《知网》的词语相似度

《知网》是一个以汉语和英语词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库 X。本文使用了《知网》提供的语义相似度的计算功能，根据论文[1]中的原理编写的词汇语义相似度计算程序实现了词语之间语义相似度的计算。

5.1.3 问题求解

由于该问题是一个 NP 完全问题[18]，本文引入模拟退火的思想，将问题求解转化为在目标函数的解空间中搜索最优解的过程。

模拟退火算法是局部搜索算法的扩展，它不同于局部搜索之处是以一定的概率选择领域中的最优值状态，理论上已经证明它是一个全局最优算法并且以概率 1 接近最优值。

基于模拟退火的词语语义倾向判定算法（SOSA 算法）首先将网络随机初始化，并设定一个高的初始“温度”T(1)。模拟退火算法能否找到全局最优解，取决于初始温度 T(1)是否足够高以及温度下降得是否足够慢，而这些正好与程序收敛时间相矛盾。为了平衡解的质量与收敛速度，我们通过实验将算法的这些参数调整为比较合适的值。然后随机地选择一个节点 $\boldsymbol{i}$，假定其现在的状态是 $\boldsymbol{w_i}=+1$，计算在这种构型下的系统总能量 $\boldsymbol{E_a}$ ，接着，再计算如果改变到候选状态，即 $\boldsymbol{w_i}=-1$ 时，对应的系统能量 $\boldsymbol{E_b}$ ，如果候选状态的能量 $\boldsymbol{E_b}<\boldsymbol{E_a}$，则接受这个状态改变；如果能量 $\boldsymbol{E_b}$ 反而更高，则以概率 $\boldsymbol{exp}\left\{-\Delta\boldsymbol{E_{ab}}/\boldsymbol{T}(\boldsymbol{k})\right\}$ 接受这个状态的改变。其中，$\Delta\boldsymbol{E_{ab}}=\boldsymbol{E_b}-\boldsymbol{E_a}$。

SOSA 算法持续多次随机轮询（选择并测试）节点，并根据以上方式进行状态改变。然后，逐渐将温度下降，重复下一轮操作。接受能量增加的候选状态的概率也逐步下降。算法继续进行，直到每个节点都被访问多次后，温度进一步下降，查询过程也重复进行。当温度非常低时，接受能量增加的状态转移的概率非常小，此时系统的行为类似贪心算法。

5.1.4 实验结果分析

本节实验使用了情感博客、电影评论和笔记本电脑三个主题的中文评论数据。

表8. 40 组褒贬基准词

褒义基准词					贬义基准词				
好	光环	得体	震动人心	巅峰	不良	弊病	痴呆	莫名其妙	粗暴
福分	感激	魅力	别具匠心	大师	功利主义	固步自封	愁	惨不忍睹	悲惨
昌盛	出色	淳朴	甘之如饴	独到	诡异	狠毒	假冒	吹毛求疵	粗鄙
绚丽	恰到好处	优秀	才华横溢	创造力	尖酸	浑浑噩噩	暴殄天物	自不量力	抱佛脚
积极	如火如荼	著名	力透纸背	逼真	薄情	保守	饱食终日	崩溃	畸形
舒服	灿烂	纯真	飞扬	青春	画蛇添足	委屈	坏	变态	失败
美好	和谐	宽容	自由自在	欢愉	呆板	游离	走火入魔	痛苦	毛病
成熟	诚实	善良	和平	文明	煽情	噱头	支离破碎	郁闷	扭曲

词语的语义倾向判断具有不确定性，表现在两个方面：首先，部分词语在不同的语用环境下具有不同的语义倾向。其次，对于同一个词，不同的人的判断也是有差异的。为减少上述因素产生的影响，本文在从文档测试集生成词语测试集时，采用多人共同标注的方法，并在构建通用领域测试集时尽量避免选择语义倾向与语用领域相关的词语。最终，本文共生成 3 个词语测试集，分别用 Set1、Set2、Set3 表示。

表 8 中列出了实验中所使用的基准词。

最终词汇实验结果如表 9 所示：

表9. 词汇实验结果

方法	HowNetPMI			PCJaccardSA			PCOverlapSA		
测试集	Set1	Set2	Set3	Set1	Set2	Set3	Set1	Set2	Set3
准确率(%)	78.6	80.7	81.2	88.9	87.8	88.2	88.4	86.9	87.4
方法	PCDiceSA			PCPMISA			HowNetSA		
测试集	Set1	Set2	Set3	Set1	Set2	Set3	Set1	Set2	Set3
准确率(%)	92.6	90.3	90.1	89.7	88.2	87.7	95	90.3	90.6

该实验证明了本文方法的有效性和实用性(详细介绍请参见[21])。

5.2 基于 Modularity 优化的通用情感词典构建

本节提出了一种新的利用函数优化进行词语语义倾向计算的方法，该方法可以自动地从字典或语料集中生成带有语义倾向的词语列表。该方法主要采用基于 modularity 优化的算法，可以实现较高的词语语义倾向计算准确率。

5.2.1 基本原理

本文采用以下的假设：具有较高相似度的词语通常具有相同的语义倾向。如前文所述，图划分方法能够更好地利用词语间的全局信息，所以本文从图划分的角度进行词语语义倾向计算。

通常，以“最小切分”为目标的目标函数需满足§5.1.1 提出的几个条件。对于该目标函数，如果将所有节点划为一类，则无疑可以使其得到极值，但这样的平凡解是没有意义的。

社区发现研究是对图划分方法的深入和扩展，其中具有代表性的方法是基于 modularity（组合性）优化的方法。modularity 是由纽曼（M.E.J.Newman）提出的，最早是作为衡量网络划分好坏的一种度量。modularity 值（也叫做 Q 值）的通常计算方法为

$$Q=\sum_{j}\left(e_{ij}-a_i^2\right),$$

其中，e_{ij} 表示社区 i 和社区 j 之间的连边占总边数的比例；

$$a_i=\sum_{j}e_{ij}$$

表示有一个端点在社区 i 中的边占总边数的比例。基于 modularity 优化的方法与图划分方法的目标是一致的，但该方法能够避免图划分方法易于陷入平凡解的弱点，因此本文采用其进行词语语义倾向计算。

5.2.2 算法基本过程

本文采用以下步骤进行词语语义倾向计算。

第一步，**构建词语相似度矩阵** 本文采用两种词语相似度计算方法构建词语相似度矩阵。第一种方法，利用《知网》提供的相似度函数；第二种方法，利用语料中词语的共现信息。

第二步，**词语语义倾向计算** 基于之前得到的词语相似度矩阵，我们以 modularity 为目标函数，按照能够使函数值极大的方式将其划分为两个不相交的子图。

具体步骤是：

– 通过词语相似度邻接矩阵构建modularity矩阵。
– 找到对应于最大特征值的特征向量，向量中每个元素对应每个待计算语义倾向的词语，将这些词语按照元素值的正负分为两类。
– 对于每类词语，首先手工确定该类中对应最大元素值的词语的语义倾向，用其作为这个类别的语义倾向。
– 持续地在两类之间交换词语，直到modularity值稳定。

返回测试集中每个词语的语义倾向。

5.2.3 实验结果分析

本小节使用教育评论、电子产品评论和股票评论三种语料，整体测试基于 modularity 优化的方法在《知网》生成的测试集以及共现测试集上的准确率。

表10. PMI[4]方法中使用的基准词

褒				贬			
词语	页面数 (单位:百万)	词语	页面数 (单位:百万)	词语	页面数 (单位:百万)	词语	页面数 (单位:百万)
好	2,400	活力	77.2	错误	214	糊涂	29.4
积极	220	舒服	69.6	不良	190	毛病	26.2
优秀	219	出色	53.7	痛苦	96.4	悲惨	24.4
漂亮	203	感激	50.5	郁闷	68.8	扭曲	18.1
高手	142	高峰	48.6	保守	44	假冒	15.1
成熟	127	善良	43.6	愁	43.1	畸形	12.8
美好	114	诚实	27	虚假	41.7	痴呆	10.5
和谐	113	宽容	23.8	变态	37.5	心酸	10.4
福分	79.7	光辉	23.4	极端	36.5	耻辱	9.99
和平	78.2	礼貌	21.5	崩溃	34.5	粗暴	8.5

本文使用中文分词软件 ICTCLAS[5]从网页中抽取词语。抽取出的词语如果也在《知网》中出现，则将其加入词语测试集 Termset1 中。如前所述，词语的语义倾向判断具有不确定性。为减少上述因素产生的影响，本文在词语测试集 Termset1 的基础上生成测试集 Termset2 和 Termset3。生成时，采用多人共同标注的方法，并尽量避免选择语义倾向与语用领域相关的词语。本文实验的另外三个词语测试集为利用语料中的词语共现信息得到，采用整篇文档作为共现窗口。在去除孤立节点（未与任何其他词语在语料中共现）后，得到三个测试集，编号分别为 4、5、6。

为验证基准词对于实验结果的影响，我们让多人各自选出一些富含语气且倾向明确的词语作为候选基准词。然后将这些词语用搜索引擎 Google 进行查询，将所有词语按照查询返回的相关页面数进行排序，并选择页面数量最多的 20 对词语作为基准词。基准词的详细信息如上面表 10 所示。

最后，整体测试基于 modularity 优化的方法在《知网》生成的测试集以及共现测试集上的准确率，结果如表 11 所示。

[4] pointwise mutual information，点间互信息（亦有译作逐点互信息），信息论或统计理论中用于度量相关度
[5] www.searchforum.org.cn

表11.《知网》HowNet 测试集上的实验结果

方法	PMI			K-L[6]			本文方法		
测试集	1	2	3	1	2	3	1	2	3
准确率	0.642	0.694	0.843	0.603	0.706	0.874	0.617	0.718	0.888
平均准确率	0.726			0.728			0.741		

表12.共现测试集上的实验结果

方法	PMI			K-L			本文方法		
测试集	4	5	6	4	5	6	4	5	6
准确率	0.457	0.456	0.446	0.592	0.583	0.576	0.618	0.598	0.596
平均准确率	0.453			0.583			0.604		

从表 11 可以看到，本文方法在测试集 2 和测试集 3 上的准确率均高于另外两种方法；在测试集 1 上，准确率低于 PMI 方法，部分原因是由于，经人工挑选后，测试集 2 和测试集 3 中的词语具有更明确的语义倾向，显示了较显著的摄取性，使得本文方法获得了较高的准确率。

从表 12 中可以看到，在三个共现测试集上本文方法的准确率均高于另外两种方法，并且本文方法在三个测试集上的准确率是稳定的，说明本文方法对于语料规模大小相对较不敏感。

5.3 基于扩展信息瓶颈的领域情感词典构建

人类的语气表达具有极强的领域相关性，因此，在实际应用中，为了获得更好的文本倾向性分析的性能，需要为每个领域建立各自的相关情感词典；而由于领域众多，由人工构建领域情感词典是不切实际的。因此，寻找到快速、实用的领域情感词典构建算法就成了极为重要的工作。我们需要解决如何利用一个已知领域（即源领域）的标注数据对另一个领域（即目标领域）进行倾向性分析，这就是跨领域倾向性分析问题。解决好此问题，才能真正有助于倾向性分析的大范围应用。目前多数方法只考虑了源领域词语与目标领域词语之间的关系，忽略了源领域文档与目标领域词语之间的关系以及目标领域词语文档之间的关系。针对该问题，本文提出一个基于信息瓶颈方法[19]的迭代增强模型，来整合源领域与目标领域的信息。

5.3.1 基本原理

本文方法基于以下假设：

1. 包含褒义词较多的文档表现为正面语气倾向；被较多正面文档包含的词语表现为褒义语义倾向，对贬义词亦是如此。
2. 尽管源领域与目标领域中情感词的分布有所差异，但两个领域之间一定存在一部分共同的部分。

基于这两个假设，才能利用源领域中与目标领域公共的那部分知识对目标领域情感词典构建进行指导。本文更进一步定义了 3 种关系，用来指导目标领域情感词典构建：

[6] 库尔贝克-莱布勒

- WDintra-Relationship：代表目标领域中的情感词与文本之间的关系。
- WWinter-Relationship：代表源领域中的情感词和目标领域情感词之间的关系。
- WDinter-Relationship：代表源领域中的文本与目标领域情感词之间的关系。

本文提出一种领域情感词典构建模型，将上述三种关系融入一个统一的框架进行考量。

5.3.2 信息瓶颈方法

信息瓶颈方法由提斯比（Naftali Tishby）等人提出[19]，其基本思想是：给定两个随机变量 X 和 Y 的联合分布 $p(x,y)$，压缩其中一个随机变量 X，同时尽量维持两个变量之间的互信息 $I(x,y)$。和著名的率失真理论相似，我们要在尽可能压缩 X 的表示长度和尽可能地保留 Y 的信息之间做出折中。每种压缩对应一种从 X 到 C 的赋值 $p(c|x)$， $p(c|x)$ 表示 X 的一个取值 x 对应 C 中一个取值 c 的概率。一般情况下，每个 x 可以对应 C 中多个甚至所有取值 c，这种情况称为软赋值；如果一个 x 对应一个 c，则称这种赋值为硬赋值。信息瓶颈方法试图找出一种最优赋值。它通过计算条件概率 $p(y|x)$ 与 $p(y|c)$ 之间的库尔贝克-莱布勒（Kullback-Leibler）距离[20]来度量 x 与 c 的距离。

$$D_{KL}\left[p(y|x)\|p(y|c)\right]=\sum_y p(y|x)\log\frac{p(y|x)}{p(y|c)} \quad (5)$$

5.3.3 将领域知识引入信息瓶颈模型

传统的利用信息瓶颈的词聚类方法在聚类过程中只考虑了词语与文本的关系。针对本文要解决的问题，我们对信息瓶颈进行扩展，用以将更多的源领域信息引入模型，来完成目标领域情感词典构建任务。令 $I(W_0,D_0)$ 代表 WDintra-Relationship， $I(W_i,W_0)$ 代表 WWinter-Relationship， $I(W_0,D_i)$ 代表 WDinter-Relationship。这样，传统信息瓶颈方法在聚类过程中的损失函数可以通过引入源领域知识而被扩展为：

$$I(D_0,W_0)-I(\hat{D}_0,\hat{W}_0)+\alpha\left[\left(I(D_i,W_0)-I(D_i,\hat{W}_0)\right)+I(W_i,W_0)-I(W_i,\hat{W}_0)\right] \quad (6)$$

据此，本文提出改进的信息瓶颈算法如下[23]：

第一步，初始化联合概率分布；

第二步，将 t (迭代步数)赋为 1；

第三步，重复迭代：

计算文本聚类，更新概率分布；
计算词聚类，更新概率分布；
t 赋为 $t+2$；
直到收敛为止。

5.3.4 实验结果分析

本小节使用酒店评论、电子产品评论以及股票评论三个领域的数据对我们提出的算法进行验证，实验结果如表 13、14 所示。

从表中可见，本文方法几乎在所有的任务上均表现出较好的性能。可能的原因在于，基准方法只考虑了源领域与目标领域词语之间的关系，忽略了其他两种关系，而本文方法由于充分利用了源领域与目标领域的信息，在通用词典及领域词典构建任务上均表现出较好的性能。

表13. 领域相关词语的分类结果

	基准方法			本文方法
	PMI	SM+SO	LE	
电子→酒店	68.4	73.5	73.2	**87.5**
电子→股票	57.8	60.6	63.1	**73.2**
酒店→电子	72.1	75.4	**76.3**	75.9
酒店→股票	73.7	76.4	78.1	82.2
股票→电子	70.6	73.3	73.4	74.1
股票→酒店	68.8	71.2	73.6	**82.8**
平均精度	68.5	71.7	72.9	**79.2**

表14. 领域无关词语的分类结果

	基准方法			本文方法
	PMI	SM+SO	LE	
电子→酒店	76.6	77.5	80.7	**88.1**
电子→股票	69.7	68.3	71.3	**73.6**
酒店→电子	74.1	76.7	**83.4**	79.7
酒店→股票	85.4	**88.0**	86.7	84.8
股票→电子	70.5	73.3	**81.3**	76.7
股票→酒店	67.9	71.2	81.8	**84.8**
平均精度	74.4	75.4	80.8	**81.2**

6 跨领域情感倾向性分析

如上节所述，很多研究者用监督分类方法解决倾向性分析问题。但是，该方法需要一个条件来保证分类的准确性：训练数据与测试数据应同分布以便测试数据可以与训练数据共享信息。然而，不同领域中的已标注数据量存在很大差异：在一些传统领域中有大量标注好的带有情感倾向性的文本，而在其它领域中却很少有已标注好的情感数据。手工标注可靠的情感数据需要大量的人工劳动，因此，我们需要解决跨领域倾向性分析问题，以实现倾向性分析的大范围应用。

本节以提高跨领域情感倾向性分析的精度为目标，从多个角度来研究该问题。

6.1 监督学习倾向性分析方法的领域移植

本小节通过分析源领域与目的领域的特征空间之间的差异与共性，提出有效的领域移植策略，消除特征空间差异对分类器性能造成的负面影响，建立可跨领域移植的倾向性分析方法[3]。

6.1.1 基本原理

如图 2 所示，源领域（Old Domain）样本分别用两个椭圆来表示，其中灰色的椭圆表示负面样本，白色的椭圆表示正面样本；目的领域（New Domain）样本也同样分别用灰白两个椭圆来表示，灰色表示负面样本，白色表示正面样本。C_{ON} 和 C_{OP} 分别是源领域负面和正面两个类别的中心向量，源领域中心线（Old Middle Line）为连接 C_{ON} 和 C_{OP} 的垂直平分线。从另一角度来看，源领域中心线实际上代表了分隔源领域负面和正面两个类别的超平面。利用源领域中心线，我们可以正确区分源领域负面和正面两个类别的样本。然而，从图中我们可以看到，对于目的领域，源领域中心线就无法正确划分正负面了，中心线以下的负面样本将会被错误地划分为正面类别，这也解释了源领域分类器应用于目的领域分类性能很差的原因。解决这个问题的一种直观的方法如图 3 所示。首先，从目的领域中挑选一些最具有领域特点的样本，如图所示。并重新训练基本分类器；然后，计算两类样本的中心 C_{SNN} 和 C_{SNP}，得到目的领域中心线（New Middle Line），此时，我们可观察到，目的领域的样本基本可以被目的领域中心线正确地分类了[26]。

具体的步骤描述如下：

(1). 基于源领域中已标注的样本训练一个分类器；
(2). 使用该分类器标注目的领域中最具有领域特点的部分样本；
(3). 利用这些样本训练出一个目的领域的分类器；
(4). 使用新的分类器标注目的领域的样本。

显然，步骤中的重点和难点是如何选择目的领域中最具有领域特点的部分样本进行标注，下一节将就这一问题进行深入论述。

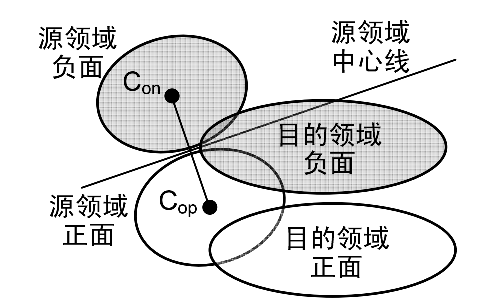


图2 源领域分类器应用于目的领域的分类性能示意图

6.1.2 样本选择方法

（1） 相似度排序方法

我们使用中心向量分类方法计算样本与正面类别的相似度 S_P，以及与负面类别的相似度 S_N，并认为：对于任意一个样本，S_P 越大，它属于正面类别的概率越大； S_N 越大，它属于负面类别的概率越大。

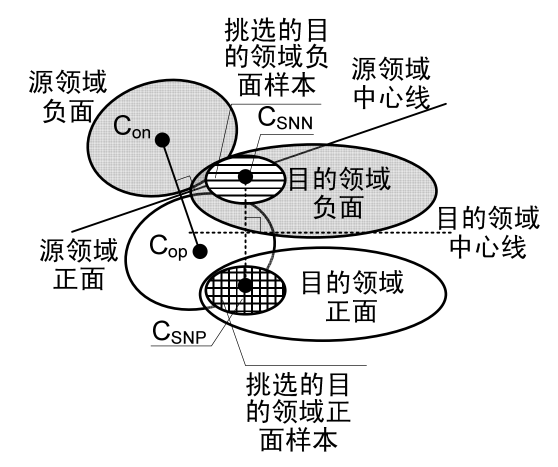


图3 领域移植后的分类器应用于目的领域的分类性能示意图

在此基础上，我们提出相似度排序（Similarity Ranking，SR)方法：即对所有样本的 S_N 进行排序，并将其中 S_N 值较大的 n/2 个样本标注为负面类别；对所有样本的 S_P 进行排序，并将其中 S_P 值较大的 n/2 个样本标注为正面类别。

（2） 相对相似度[7]方法

然而，如果网络评论的长度相差太大，这种方法就不再有效了，因为长度很长的评论通常会具有较大的 S_N 或 S_P 。另外，即使源领域和目的领域的评论长度相仿，其特征空间的较大差异也将导致 S_N 或 S_P 的较大差别。

为解决这个问题，我们将计算得到的相似度进行了规范化，从而弥补了因长度或特征空间变化所造成的不利影响，这就是相对相似度方法的基本思想。我们定义负面相对相似度（S_{RN}）和正面相对相似度（S_{RP}）如下：

$$S_{RN} = \frac{S_N}{(S_N + S_P)/2} \quad (7)$$

$$S_{RP} = \frac{S_P}{(S_N + S_P)/2} \quad (8)$$

可以认为：对于任意一个样本， S_{RP} 越大，它属于正面类别的概率越大； S_{RN} 越大，它

[7] Relative Similarity Ranking，RSR

属于负面类别的概率越大。据此，我们可以得到相对相似度排序方法：即对所有样本的 S_{RN} 进行排序，并将较大的 n/2 个样本标注为负面类别；对所有样本的 S_{RP} 进行排序，并将较大的 n/2 个样本标注为正面类别。

相对相似度排序方法的具体实现步骤如下：

(1). 计算样本的相似度 S_P 和 S_N；
(2). 采用公式（7）和公式（8）计算 S_{RP} 和 S_{RN}；
(3). 对 S_{RP} 和 S_{RN} 分别进行排序；
(4). 将 S_{RN} 较大的 n/2 个样本标注为负面类别，将 S_{RP} 较大的 n/2 个样本标注为正面类别，其中，n 表示目的领域中预先设定的一定比例（Ratio）样本的数量。

6.1.3 实验结果分析

为了验证本章提出的领域移植方法的有效性，我们采用了三个领域的数据：电脑评论，教育评论和房产评论。

表15. 领域移植方法的实验

	中心向量法	直推式向量支持机[8] 方法(基准)	领域移植方法	
			相似度排序	相对相似度排序
电脑→教育	0.7993	0.6887	0.6966	**0.8530**
电脑→房产	0.4540	**0.8960**	0.8320	0.8440
教育→电脑	0.5053	0.6509	0.7751	**0.8051**
教育→房产	0.5120	0.6100	**0.8280**	0.7200
房产→电脑	0.7387	0.7815	0.8094	**0.8993**
房产→教育	0.5781	0.6840	0.7109	**0.8214**
平均精度	0.5979	0.7185	0.7753	**0.8238**

实验中采用了本章所提出的两种领域移植方法（相似度排序方法和相对相似度排序方法）。目的领域的数据被平均地分为未标注集和测试集，未标注集的挑选比例（Ratio）为 0.4。

从表 15 的结果可以看出，相对相似度排序方法大大提高了目的领域中分类器的性能，精度提高的幅度如此显著，显示了基于相对相似度排序方法的领域移植方法具有很好的鲁棒性和有效性。

尽管相似度排序方法较为简单和直接，其性能也不错，相比相对相似度排序方法平均精度低大约 5%，但相比中心向量法却高出 18% 。在“教育→房产”领域移植实验中，其性能甚至优于相对相似度排序方法。

在领域移植实验中，相比中心向量法，直推式向量支持机方法的性能较好。除“电脑→教育”实验外，直推式向量支持机方法的精度均比中心向量法高很多，平均精度高 12% 。然而，除“电脑→房产”实验中，直推式向量支持机方法的精度都比相似度排序方法和相对相似度排序方法要低很多，这说明本小节提出的领域移植方法优于直推式向量支持机方法。

6.2 基于贝叶斯学习的情感移植模型

本小节的研究针对如何最大限度利用源领域和目标领域数据。为平衡源领域数据，我们提出频繁共现熵，挑选出两个领域中频繁出现且有相似发生概率的通用情感特征；为获得目标领域信息，我们提出自适应朴素贝叶斯算法，这是朴素贝叶斯分类器的加权移植版本[21]。

[8] TSVM，transductive support vector machine

6.2.1 算法描述

贝叶斯算法是一个非常有效的监督学习方法，在情感领域也表现出了不错的性能。但是，贝叶斯算法受到词空间差异的影响。

我们的基本思想是寻找领域之间的通用情感词，并把通用情感词作为源领域通向目标领域的一座桥梁。在训练过程中，我们逐步加大目标领域的权重，使分类器模型与目标领域达到最佳匹配。可见，我们的算法既利用了源领域的部分可用信息，又充分吸收了目标领域的全部信息。

具体来说，我们提出一个算法，首先用频繁共现熵挑选出两个领域中频繁出现且有相似发生概率的共有特征，然后用自适应朴素贝叶斯算法为目标领域训练一个分类器。

6.2.2 频繁共现熵

要进行倾向性分析，可以利用与目标领域相关的情感词和通用情感词共同为目标领域训练一个情感分类器。然而目标领域中很难获得大量标注实例，也就很难获得大量与目标领域相关的情感词，因此我们只能使用通用情感词作为源领域与目标领域的桥梁。

为获得通用情感词，本文提出一个频繁共现熵算法。通用情感特征符合两个准则：（1）两个领域频繁出现；（2）有相似的出现概率。为满足这两个准则，我们提出如下公式：

$$f_w = log\left(\frac{P_o(w)\cdot P_n(w)}{\left|P_o(w)-P_n(w)\right|+\beta}\right) \tag{9}$$

其中，$P_o(w), P_n(w)$ 表示特征 w 在源领域与目标领域的出现概率；β 的引入是为了防止出现分母为 0 的情况，在我们的方法中 $\beta = 0.0001$。

6.2.3 自适应朴素贝叶斯算法

本小节我们将基于期望最大化[9]的朴素贝叶斯方法（记为EMNB）用于跨领域学习。原则上说，EMNB要求标注数据和未标注数据服从同分布。很明显我们的跨领域学习问题不满足这一要求。然而，如果使用频繁共现熵方法挑选通用特征并且只使用这些特征初始化朴素贝叶斯模型进行期望最大化迭代，即可解决此问题。另一个问题是：只使用通用特征不足以准确预测目标领域标签。为解决此问题，我们提出一个新的加权EMNB分类器：随着迭代的进行，逐渐增加目标域数据的权重，减少源领域数据的权重，同时使用所有目标领域特征，从而极大增强分类器对目标域的预测能力。

期望最大化算法迭代两步（E步和M步）找到 $l(\theta|D)$ (详细介绍请参见[21])的局部最大参数：

E步：

$$P(c_k|d_i) \propto P(c_k)\prod_{t\in|V|}\left(P(w_t|c_k)\right)^{N_{t,i}} \tag{10}$$

M步：

$$P(c_k) = \frac{(1-\lambda)\cdot\sum_{i\in D^o}P(c_k|d_i)+\lambda\cdot\sum_{i\in D^n}P(c_k|d_i)}{(1-\lambda)\cdot|D^o|+\lambda\cdot|D^n|} \tag{11}$$

$$P(w_t|c_k) = \frac{(1-\lambda)\cdot(\eta_t^o\cdot N_{t,k}^o)+\lambda\cdot(N_{t,k}^n)+1}{(1-\lambda)\cdot\sum_{t=1}^{|V|}(\eta_t^o\cdot N_{t,k}^o)+\lambda\cdot\sum_{t=1}^{|V|}(N_{t,k}^n)+|V|} \tag{12}$$

---
[9] Expectation-Maximization，EM

详见文章[21]。

6.2.4 实验结果分析

为验证本算法，分别使用教育评论、股票评论、电脑评论三个数据集。首先验证频繁共现熵算法，以下为股票评论和电脑评论间的前 40 个通用特征：

表16. 股票评论和电脑评论间的前 40 个通用特征

不错	不明	独立	寒	快	慢	能够	凸显	庸	有效
不好	不能	精彩	喜	快速	美	骗	无法	优	正常
不及	不足	好	特色	垄断	难	伤	吸引	优势	正面
不利于	差	好处	困难	落后	难关	失	虚	优秀	阻

然后验证本方法总体性能（结果见表 17）。

由上述实验结果可见，本节提出的基于贝叶斯学习的情感移植模型可以挑选出很好的通用特征，并能大幅度促进跨领域情感倾向性分析的性能，是一个实用的算法。

表17. 不同方法的性能[10]

	NB		EMNB		NBTC[11]		本文方法	
	MicroF1	MacroF1	MicroF1	MacroF1	MicroF1	MacroF1	MicroF1	MacroF1
教育→股票	0.6704	0.4553	0.6628	0.4266	0.6743	0.4659	0.7669	0.7109
教育→电脑	0.5085	0.4696	0.4175	0.3118	0.6059	0.5918	0.8854	0.8814
股票→教育	0.6824	0.5867	0.6962	0.6056	0.8303	0.8080	0.9171	0.9119
股票→电脑	0.5053	0.5025	0.5192	0.5169	0.5128	0.5103	0.7901	0.7652
电脑→股票	0.6580	0.4148	0.6552	0.4036	0.6580	0.4148	0.6962	0.5942
电脑→教育	0.6114	0.4105	0.6074	0.4003	0.6114	0.4105	0.9013	0.8920
平均精度	0.6060	0.4732	0.5930	0.4441	0.6488	0.5336	0.8262	0.7926

6.3 基于图排序模型的跨领域倾向性分析算法

本小节提出将文本的情感倾向性与图排序算法结合起来进行跨领域倾向性分析的算法，本算法在图排序算法基础上，利用训练域文本的准确标签与测试域文本的伪标签来迭代进行倾向性分析[22]。

6.3.1 算法描述

图排序算法（如PageRank[16]）的思想是：在一个图中，与重要结点紧密相联的结点也很重要。该算法已成功应用于很多领域。基于图排序思想，我们认为如果一个文本与一些具有支持（反对）态度的文本紧密联系，则它也很可能持支持（反对）态度，这也是邻域学习思想。

因此，我们将训练集和测试集看作一个图，里面的每一个文本为图中的一个结点。给每一个结点一个表示其情感类别的分数，称其为情感分。本文提出的算法将文本情感类别间的关系与图排序（graph-ranking）算法结合起来。对于每一个待标注文本，算法通过其在训练域和测试域的邻域来计算它的情感分，并用一个统一的公式进行迭代。当算法收敛时，得到

[10] MicroF1 表示微平均，MacroF1 表示宏平均，详细介绍请参见[21]

[11] Naive Bayes Transfer Classifier，朴素贝叶斯迁移分类器

待标注文本的最终情感分。如果一个结点的情感分在-1 到 0 之间，表示这个结点所代表的文本是持反对态度，情感分越接近于-1，此文本反对态度越强；如果一个结点的情感分在 0 到 1 之间，表示这个结点所代表的文本是持支持态度，情感分越接近于 1，此文本于支持态度越强。

6.3.2 基于图排序模型的跨领域倾向性分析算法

6.3.2.1 算法初始化

第一步，本算法需要为训练集与测试集中每一个文本的情感分赋初始值，得到初始情感分向量 $\boldsymbol{S}^0=\left\{s_1^{(0)},\cdots,s_n^{(0)},s_{n+1}^{(0)},\cdots,s_{n+m}^{(0)}\right\}$。对于测试集中的文本，使用典型的文本分类算法中的任一种分类器，用训练集训练，对测试集分类得到一个伪标签（此时的准确度通常很低）。对于每一个文本，如果它分配到的标签是“反对”，则将它的情感分赋为-1；如果它分配到的标签是“支持”，则将它的情感分赋为 1。

第二步，为保证最终程序的收敛性，将测试集对应的情感分初始值 $s_i^{(0)}\left(i=1,\cdots,n\right)$ 归一化，使得正的情感分的和为 1，负的情感分的和为-1。同样，将训练集对应的情感分初始值 $s_j^{(0)}\left(j=n+1,\cdots,n+m\right)$ 归一化。

6.3.2.2 情感分计算策略

得到初始情感分向量 $\boldsymbol{S}^0$ 后，即可利用训练域的准确情感分和测试域的伪情感分来迭代计算测试集的最终情感分。

首先，利用训练集的准确情感分来计算测试集的情感分。建立一个图模型，结点表示源领域标注文本集 $\boldsymbol{D}^L$ 和目标领域未标注文本集 $\boldsymbol{D}^U$ 中的文本，边表示文本间的内容相似度。如果两个文本间内容相似度为 0，则图中两点间无边。如果不为 0，则图中两点间有边，且边的权重即为此两内容之间的相似度。内容相似度有很多方法求出，此处用余弦相似度来计算。我们使用一个联接矩阵来表示 $\boldsymbol{D}^U$ 和 $\boldsymbol{D}^L$ 间的相似矩阵。为保证算法收敛，将联接矩阵归一化，使得归一化后矩阵中每一行的和为 1。为了找出与一个文本最相似的文本集（此处设此文本集大小为 K），我们对归一化后矩阵的每一行进行降序排列，因此对于 $d_i\in\boldsymbol{D}^U\left(i=1,\cdots,n\right)$,得到它在训练域中的 K 个邻居。

其次，利用测试集的“伪”情感分来计算测试集的情感分。这与利用训练集的方法类似。

6.3.2.3 算法迭代过程

本算法要同时利用训练域和测试域的信息来对测试域的文本进行标注，因此综合利用训练集中邻域的情感分和测试集中邻域的伪情感分，得到迭代计算测试数据集的情感分的公式如下所示：

$$s_i^{(k)}=\alpha\sum_{j\in N_i}\left(\hat{U}_{ij}\times s_j^{(k-1)}\right)+\beta\sum_{h\in M_i}\left(\hat{V}_{ih}\times s_h^{(k-1)}\right),\quad i=1,\cdots,n \qquad (13)$$

其中 $\alpha+\beta=1$, α 和 β 分别表示训练域和测试域对最终情感分的贡献大小。为保证算法收敛，算法每迭代一次都需要将 $\boldsymbol{S}$ 归一化，使得正的情感分之和为 1，负的情感分之和为-1。迭代计算情感分 $\boldsymbol{S}$ 并归一化，直到算法收敛为止。

6.3.3 实验结果分析

为验证本小节提出的算法的性能，本节针对电子评论、财经评论以及酒店评论进行实验，并将该算法与其他典型算法进行比较分析。本节我们用支持向量机来初始化本文提出的算法中的情感分。

表18. 跨领域倾向性分析时不同算法性能比较

	LibSVM[12]	SCL[13]	本文算法
电子→财经	0.6478	**0. 7507**	0.7304
电子→酒店	0.7522	**0. 7750**	0.7543
财经→酒店	0.6957	**0. 7683**	0.7457
财经→电子	0.6696	0. 8340	**0.8435**
酒店→财经	0.5978	0. 6571	**0.7848**
酒店→电子	0.6413	0. 7270	**0.8609**
平均精度	0.6674	0. 7520	**0.7866**

由表 18 可以看出，基于图排序的跨领域倾向性分析算法大幅度地提高了跨领域倾向性分析的精度。其中第 2 列是LibSVM的精度，第 4 列为用LibSVM初始化后本算法的精度，对比可见，我们算法的精度均高于LibSVM的精度，平均精度提高了 11.9%。精度上如此大幅度的提高表明我们的算法对于跨领域倾向性分析问题非常有效。

7 总结和展望

文本情感倾向性分析的相关研究得到国家自然科学基金、国家 863 计划等多方面的项目资助。我们以提高文本倾向性分析精度为目标，分别从整篇文本的倾向性分析、领域情感词典构建及跨领域情感倾向性分析三方面提出相应的解决方法，从而通过不同角度提高文本倾向性分析的精度。在深入开展技术研究的同时，我们已经陆续开发出实用性系统，帮助用户准确迅速地判断文本的情感倾向性。在下一步的工作中，将在以上各方面进行深入研究，进一步促进倾向性分析的大范围应用。

参考文献：

[1] 杜伟夫,谭松波,云晓春,程学旗.一种新的情感词汇语义倾向计算方法. 计算机研究与发展. 2009,46(10): 1713-1720

[2] 唐慧丰，谭松波，程学旗. 基于监督学习的中文情感分类技术比较研究. 中文信息学报. 2007,21(6):88-94

[3] 徐俊，蔡莲红. 面向情感转换的层次化韵律分析与建模. 清华大学学报（自然科学版）. 2009，49(S1): 1274-1277

[4] B. Liu, M. Hu, J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. In: Proc of the 14th international conference on World Wide Web, Chiba, Japan, 2005: 1367-1373

[5] 宋鸿彦，刘军，姚天昉，等. 汉语意见型主观性文本标注语料库的构建. 中文信息学报. 2009, 23(2): 123-128

[6] 朱嫣岚，闵锦，周雅倩，等. 基于 HowNet 的词汇语义倾向计算. 中文信息学报, 2006, 20 (1) : 14-20

[7] H. Tang, S. Tan, and X. Cheng. A Survey on Sentiment Detection of Reviews. Expert Systems with Applications. 2009, 36(7): 10760-10773.

[8] 王根，赵军. 基于多重冗余标记 CRFs 的句子情感分析研究. 中文信息学报. 2007, 21(05) : 51-56

[12] 台湾大学林智仁(Lin Chih-Jen)等开发设计的一个简单、易于使用和快速有效的 SVM 模式识别与回归的软件包

[13] structural correspondence learning，结构对应学习算法

[9] P. D. Turney, M. L. Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Transactions on Information Systems, 2003, 21 (4): 315-346

[10] H. Chen, M. Lin, Y. Wei. Novel Association Measures Using Web Search with Double Checking. In: Proc of the COLING/ACL, 2006: 1009-1016

[11] B. Pang, L. Lee, S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In: Proc of EMNLP, Philadelphia, USA, 2002: 79-86

[12] H. Cui, V. Mittal, and M. Datar. Comparative experiments on sentiment classification for online product reviews. In: Proc of AAAI, Boston, USA, 2006:1265-1270

[13] M. Hu, B. Liu. Mining and Summarizing Customer Reviews. In: Proc of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, 2004: 168-177

[14] 胡熠,陆汝占,李学宁,段建勇，陈玉泉. 基于语言建模的文本情感分类研究.计算机研究与发展. 2007, 44(9): 1469-1475

[15] 徐琳宏,林鸿飞,杨志豪.基于语义理解的文本倾向性识别机制.中文信息学报,2007,21(1): 96-100

[16] S. Brin, L. Page, R. Motwani, and T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web，Tech. Rep. 1999-66, Stanford Digital Libraries

[17] 邵艳秋，韩纪庆，王卓然，刘挺. 韵律参数和频谱包络修改相结合的情感语音合成技术研究. 信号处理, 2007, 23(4): 526-530

[18] Christos H P. Computational complexity. New York: Addision-Wesley Publishing Company,1994: 50-60

[19] N. Tishby, F. C. Pereira, W. Bialek. The Information Bottleneck Method. In: Proc of 37th Allerton Conferenct on Communication and Computation, 1999

[20] T. Cover, J. Thomas. Elements of Information Theory. NewYork: Wiley-Interscience, 1991

[21] Songbo Tan, Xueqi Cheng, Yuefen Wang et al. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. In: Proc of 31st European Conference on Information Retrieval, Springer Berlin, Heidelberg, 2009: 337-349

[22] Q. Wu, S. Tan, H. Zhai, et al. SentiRank: Cross-Domain Graph Ranking for Sentiment Classification. In: Proc of Web Intelligence, 2009

[23] W. Du, S. Tan, X. Cheng and X. Yun. Adapting Information Bottleneck Method for Automatic Construction of Domain-oriented Sentiment Lexicon. In Proceedings of WSDM 2010.

[24] Weifu Du, Songbo Tan: Building domain-oriented sentiment lexicon by improved information bottleneck. CIKM 2009: 1749-1752

[25] Weifu Du, Songbo Tan: An Iterative Reinforcement Approach for Fine-Grained Opinion Mining. NAACL 2009: 486-493

[26] S. Tan, G. Wu, H. Tang and X. Cheng. A novel scheme for domain-transfer problem in the context of sentiment analysis. In Proceedings of CIKM 2007.

作者简介：

吴 琼：中国科学院计算技术研究所 网络重点实验室 博士生
谭松波：中国科学院计算技术研究所 网络重点实验室 副研究员
Email: tansongbo@software.ict.ac.cn
程学旗：中国科学院计算技术研究所 网络重点实验室 研究员